An evidential pixel-based face blurring approach

Pauline Minary[1,2], Frédéric Pichon[1], David Mercier[1], Eric Lefevre[1], and Benjamin Droit[2]

[1]Univ. Artois, EA 3926,
Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
Béthune, F-62400, France.
{frederic.pichon,david.mercier,eric.lefevre}@univ-artois.fr
[2]SNCF Réseau,
Département des Télécommunications,
La Plaine Saint Denis, France.
{pauline.minary,benjamin.droit}@reseau.sncf.fr

Abstract. Blurring faces on images may be required for anonymity reasons. This may be achieved using face detectors that return boxes potentially containing faces. The most direct way to exploit these detectors is to combine them in order to obtain a more efficient face detection system, producing more accurate boxes. However, contrary to detection, blurring is actually a decision problem situated rather at the pixel level than the box level. Accordingly, we propose in this paper a face blurring system based on face detectors, which operates at the pixel-level. First, for each pixel, detector outputs are converted into a common representation known as belief function using a calibration procedure. Then, calibrated outputs are combined using Dempster's rule. This pixel-based approach does not have some shortcomings of a state-of-the-art box-based approach, and shows better performances on a classical face dataset.

Keywords: Belief functions, Information fusion, Image processing, Evidential calibration, Face blurring.

1 Introduction

Blurring faces on images may be required for anonymity reasons. Due to the generally large amount of images to process but also the necessity for good performances (in particular, avoiding missed faces), one must resort to semi-automatic blurring systems – typically, a human operator correcting the outputs of an automatic face detection system.

Face detection can be performed using single detectors [6, 10], yet since detectors are generally complementary, *i.e.*, they do not detect only the same faces, using multiple detectors is a means to improve overall performance. Within this scope, Faux [3] proposed a face detection system, which consists in combining outputs of the face detector proposed in [10] and a skin colour detector. This step of combination is conducted within a framework for reasoning under uncertainty called evidence theory [8, 9]. However, it does not use all available information.

Indeed, for a given image, face detectors such as [10] provide a set of bounding boxes corresponding to the assumed positions of the faces, but they provide as well for each of these boxes a confidence score.

In the context of pedestrian detection, Xu *et al.* [11] recently proposed an evidential approach, which uses these confidence scores. Specifically, multiple detectors are used in Xu *et al.* [11], and to be able to combine them, the scores they produce are transposed into a common representation; this latter procedure is called calibration [7]. Of particular interest is that Xu *et al.* [12] subsequently refined this calibration procedure, in order to account explicitly for uncertainties inherent to such process.

Now, although face blurring may be achieved using simply the bounding boxes outputted by a face detection system, we may remark that it is not exactly the same problem as face detection. Indeed, face blurring amounts merely to deciding whether a given pixel belongs to a face, whereas face detection amounts to determining whether a given set of pixels corresponds to the same face. In other words, the richer box-based information provided by detection systems is not strictly necessary for blurring. This remark opens the path for a different approach to reasoning about blurring, which may then be situated at the pixel-level rather than box-level. In particular, face detectors may still be used but their outputs need not be combined so as to produce boxes as is the case in face detection.

Accordingly, we propose in this paper a face blurring system based on face detectors, which operates at the pixel-level. First, for each pixel, detector outputs are calibrated using Xu *et al.* procedure [12]. Then, calibrated outputs are combined using Dempster's rule [8]. We may already remark that this approach does not have some shortcomings of box-based methods, as will be shown later.

This paper is organized as follows. First, Section 2 recalls necessary background on evidence theory and calibration. Then, Section 3 exposes what may be considered presently as one of the best available blurring system based on multiple detectors, that is, an evidential system relying on face detection performed using Xu *et al.* detection approach [11], applied to faces rather than pedestrians and improved using Xu *et al.* calibration [12]. Our proposed pixel-based system is then detailed in Section 4. An experiment comparing these two approaches is reported in Section 5, before concluding in Section 6.

2 Evidence theory and calibration: necessary background

2.1 Evidence theory

The theory of evidence is a framework for reasoning under uncertainty. Let Ω be a finite set called the frame of discernment, which contains all the possible answers to a given question of interest Q. In this theory, uncertainty with respect to the answer to Q is represented using a *Mass Function* (MF) defined as a mapping $m^{\Omega} : 2^{\Omega} \to [0,1]$ that satisfies $\sum_{A \subseteq \Omega} m^{\Omega}(A) = 1$ and $m^{\Omega}(\emptyset) = 0$. The quantity $m^{\Omega}(A)$ corresponds to the share of belief that supports the claim that the answer is contained in $A \subseteq \Omega$ and nothing more specific.

Given two independent MFs m_1^Ω and m_2^Ω about the answer to Q, it is possible to combine them using Dempster's rule of combination. The result of this combination is a MF $m_{1\oplus 2}^\Omega$ defined by

$$m_{1\oplus 2}^\Omega(A) = \frac{1}{1-\kappa} \sum_{B\cap C=A} m_1^\Omega(B) m_2^\Omega(C), \tag{1}$$

for all $A \subseteq \Omega$, where $\kappa = \sum_{B\cap C=\emptyset} m_1^\Omega(B) m_2^\Omega(C)$.

Different decision strategies exist to make a decision about the true answer to Q, given a MF m^Ω on this answer [1]. In particular, the answer having the smallest so-called *upper expected cost* may be selected. The upper expected cost $R^*(\omega)$ of some answer $\omega \in \Omega$ is defined as

$$R^*(\omega) = \sum_{A\subseteq\Omega} m^\Omega(A) \max_{\omega'\in A} c(\omega, \omega'), \tag{2}$$

where $c(\omega, \omega')$ is the cost of deciding ω when the true answer is ω'.

2.2 Evidential calibration of binary classifiers

A binary classifier, *e.g.*, a detector, may return a score associated to its classification decision, which is a valuable information because it provides an indication on how confident the classifier is. The range of these scores differs depending on the features and the type of the classification algorithm used. Thus, transposing scores in a common representation is essential in a context of multi-detectors.

This step, called calibration of a classifier, relies on a training set $\mathcal{L}_{cal} = \{(S_1, Y_1), ..., (S_n, Y_n)\}$, with S_i the score provided by the classifier for the i^{th} sample and $Y_i \in \Omega = \{0, 1\}$ its associated true label. Given a new score S, the purpose of calibration is to estimate the posterior probability distribution $p_S^\Omega = p^\Omega(\cdot|S)$ using $\mathcal{L}_{cal}$.

Yet, certain score values may be less present than others in $\mathcal{L}_{cal}$, thus some estimated probabilities may be less accurate than others. To address this issue, Xu *et al.* [12] proposed several evidential extensions of probabilistic calibration methods. Accordingly, given a new score S, any of Xu *et al.* [12] evidential calibration procedures yields a MF m_S^Ω (rather than a probability distribution p_S^Ω) accounting explicitly for uncertainties in the calibration process.

Among the evidential calibration procedures studied in [12], the likelihood-based logistic regression presents overall better performances than other calibrations. Thus, this will be the calibration used in this paper[1].

3 An evidential box-based face detection approach

Face blurring may be achieved using simply the boxes outputted by a face detection system. In this section, we present such a system, which may be considered

[1] Due to lack of space, we must refrain from recalling the definition of m_S^Ω obtained under this calibration. We refer the interested reader to [12].

as state-of-the-art with respect to face detection. In a nutshell, it is merely Xu *et al.* [11] evidential box-based detection approach applied to faces rather than pedestrians and whose calibration step has been replaced by the evidential likelihood-based logistic regression calibration procedure proposed in [12].

3.1 Xu *et al.* [11] box-based detection approach applied to faces

Let us consider a given image and assume that J face detectors are run on this image. Formally, each detector D_j, $j = 1, ..., J$, provides N_j couples $(B_{i,j}, S_{i,j})$, where $B_{i,j}$ denotes the i^{th} box, $i = 1, ..., N_j$, returned by the j^{th} detector and $S_{i,j}$ is the confidence score associated to this box.

Through a calibration procedure, which will be described in Section 3.2, score $S_{i,j}$ is transformed into a MF $m^{\mathcal{B}_{i,j}}$ defined over the frame $\mathcal{B}_{i,j} = \{0, 1\}$, where 1 (resp. 0) means that there is a face (resp. no face) in box $B_{i,j}$.

Then, using a clustering procedure detailed in Section 3.3, all the boxes $B_{i,j}$ returned by the J detectors for the considered image, are grouped into K clusters C_k, $k = 1, ..., K$, each of these clusters being represented by a single box B_k.

In addition, for each box $B_{i,j} \in C_k$, its associated MF $m^{\mathcal{B}_{i,j}}$ is assumed to represent a piece of evidence regarding the presence of a face in B_k, that is, $m^{\mathcal{B}_{i,j}}$ is converted into a MF $m^{\mathcal{B}_k}_{i,j}$ on $\mathcal{B}_k = \{0, 1\}$ defined by $m^{\mathcal{B}_k}_{i,j}(A) = m^{\mathcal{B}_{i,j}}(A)$, for all $A \subseteq \{0, 1\}$. These pieces of evidence are then combined using Dempster's rule; this can be done as the sources are considered to be independent and reliable. More complex combination schemes are also considered in [11]. However, only Dempster's rule, which besides presents good performance in [11], is considered here. The combination results in a MF $m^{\mathcal{B}_k}$ representing the overall system uncertainty with respect to the presence of a face in B_k.

3.2 Box-based score calibration for a detector

In order to transform the score $S_{i,j}$ associated to a box $B_{i,j}$ into a MF $m^{\mathcal{B}_{i,j}}$, detector D_j needs to be calibrated. In particular, the evidential likelihood-based logistic regression calibration procedure [12] may be used instead of the cruder procedures used in [11]. As recalled in Section 2.2, such procedures require a training set, which we denote by $\mathcal{L}_{cal,j}$. We recall below how $\mathcal{L}_{cal,j}$ is built.

Assume that L images are available. Besides, the positions of the faces really present in each of these images are known in the form of bounding boxes. Formally, this means that for a given image ℓ, a set of M^ℓ boxes $G^\ell_r, r = 1, ..., M^\ell$, is available, with G^ℓ_r the r^{th} bounding (ground truth) box on image ℓ.

Furthermore, detector D_j to be calibrated is run on each of these images, yielding N^ℓ_j couples $(B^\ell_{t,j}, S^\ell_{t,j})$ for each image ℓ, where $B^\ell_{t,j}$ denotes the t^{th} box, $t = 1, ..., N^\ell_j$, returned on image ℓ by detector D_j and $S^\ell_{t,j}$ is the confidence score associated to this box.

From these data, training set $\mathcal{L}_{cal,j}$ is defined as the set of couples $(S^\ell_{t,j}, YB^\ell_{t,j})$, $\ell = 1, ..., L$, and $t = 1, ..., N^\ell_j$, with $YB^\ell_{t,j} \in \{0, 1\}$ the label obtained by evaluating whether box $B^{\ell}_{t,j}$ "matches" some face in image ℓ, *i.e.*,

$$YB^{\ell}_{t,j} = \begin{cases} 1 \text{ if } \exists G^{\ell}_r, r = 1, ..., M^{\ell}, \text{ such that } ov(G^{\ell}_r, B^{\ell}_{t,j}) \geq \lambda, \\ 0 \text{ otherwise,} \end{cases}$$

where λ is some threshold in $(0,1)$ and $ov(G^{\ell}_r, B^{\ell}_{t,j})$ is a measure of the overlap between boxes G^{ℓ}_r and $B^{\ell}_{t,j}$ [2]. It is defined by

$$ov(B_1, B_2) = \frac{area(B_1 \cap B_2)}{area(B_1 \cup B_2)}, \tag{3}$$

for any two boxes B_1 and B_2. Informally, $\mathcal{L}_{cal,j}$ stores the scores associated to all the boxes returned by detector D_j on images where the positions of faces are known, and records for each score whether its associated box is a true or false positive. The MF $m^{\mathcal{B}_{i,j}}$ associated to a new score $S_{i,j}$ and obtained from calibration relying on $\mathcal{L}_{cal,j}$ represents thus uncertainty toward box $B_{i,j}$ containing a face.

3.3 Clustering of boxes

As several detectors are used, some boxes may be located in the same area of an image, which means that different boxes assume that there is a face in this particular area. The step of clustering allows one to group those boxes and to retain only one per cluster. A greedy approach is used in [11]: the procedure starts by selecting the box $B_{i,j}$ with the highest mass of belief on the face hypothesis and this box is considered as the representative of the first cluster. Then, for each box $B_{u,v}$, $\forall (u,v) \neq (i,j)$, such that the overlap $ov(B_{i,j}, B_{u,v})$ is above the threshold λ, the box $B_{u,v}$ is grouped into the same cluster as $B_{i,j}$, and is then no longer considered. Among the remaining boxes, the box $B_{i,j}$ with the highest $m^{\mathcal{B}_{i,j}}(\{1\})$ is selected as representative of the next cluster, and the procedure is repeated until all the boxes are clustered.

4 Proposed approach

As explained in Section 1, for the purpose of blurring, it seems interesting to work at the pixel level rather than box level. This section gives the full particulars of our proposed pixel-based system.

4.1 Overview of the approach

To each pixel $p_{x,y}$ in an image, we associate a frame of discernment $\mathcal{P}_{x,y} = \{0, 1\}$, where x and y are the coordinates of the pixel in the image and 1 (resp. 0) means that there is a face (resp. no face) in pixel $p_{x,y}$.

The inputs of our approach are the same as for the box-based approach but are treated differently. In particular, if pixel $p_{x,y}$ belongs to a box $B_{i,j}$, the score $S_{i,j}$ associated to this box $B_{i,j}$ is "transferred" to $p_{x,y}$ and then using the evidential likelihood-based logistic regression calibration procedure together with a training set $\mathcal{L}_{calP,j}$ defined in Section 4.2, this score is transformed into a MF $m_{i,j}^{\mathcal{P}_{x,y}}$. If pixel $p_{x,y}$ does not belong to any of the boxes returned by detector D_j, we take this into account *via* a MF denoted $m_{*,j}^{\mathcal{P}_{x,y}}$ and defined in Section 4.2.

Eventually, we then obtain for pixel $p_{x,y}$ several MFs on $\mathcal{P}_{x,y}$, which we combine by Dempster's rule, resulting in a MF $m^{\mathcal{P}_{x,y}}$ representing the overall system uncertainty with respect to the presence of a face in $p_{x,y}$.

This approach has in theory a high complexity. However, since we have $m_{*,j}^{\mathcal{P}_{x,y}}(A) = m_{*,j}^{\mathcal{P}_{x',y'}}(A)$, for all $A \subseteq \{0,1\}$ and $x' \neq x$ or $y' \neq y$, *i.e.*, any two pixels that do not belong to a box of D_j are associated MFs with the same definitions, then pixels that do not belong to any of the returned boxes by the detectors have the same resulting MF. Hence, since this latter case happen often in practice, this allows us to have a common processing for a very large number of pixels, which considerably reduces the complexity.

Let us finally remark that this approach presents several advantages over the one of Section 3: first, as will be seen in Section 4.2, our calibration step avoids the use of the parameter λ, whose value needs to be fixed either *a priori* (but then it is arguably arbitrary) or empirically; second, our approach avoids the use of clustering, which also involves the parameter λ and that may behave non optimally in a multi-object situation, especially when they are close to each other, which may be the case with faces.

4.2 Pixel-based score calibration for a detector

Let us describe the set $\mathcal{L}_{calP,j}$ underlying the transformation using calibration of a score $S_{i,j}$ associated to a pixel $p_{x,y}$ by a detector D_j, into a MF $m_{i,j}^{\mathcal{P}_{x,y}}$.

For a given image ℓ, each couple $(B_{t,j}^{\ell}, S_{t,j}^{\ell})$ introduced in Section 3.2 yields, *via* "transfer", $\left|B_{t,j}^{\ell}\right|$ couples $(p_{d,t,j}^{\ell}, S_{t,j}^{\ell})$, with $d = 1, \ldots, \left|B_{t,j}^{\ell}\right|$, and $\left|B_{t,j}^{\ell}\right|$ the number of pixels in box $B_{t,j}^{\ell}$, and where $p_{d,t,j}^{\ell}$ denotes the pixel in d^{th} position in box $B_{t,j}^{\ell}$.

From these data, we define $\mathcal{L}_{calP,j}$ as the set of couples $(S_{t,j}^{\ell}, YP_{d,t,j}^{\ell})$, with $\ell = 1, \ldots, L$, $t = 1, \ldots, N_j^{\ell}$, and $d = 1, \ldots, \left|B_{t,j}^{\ell}\right|$, with $YP_{d,t,j}^{\ell} \in \{0,1\}$ the label simply obtained by checking whether pixel $p_{d,t,j}^{\ell}$ belongs to some ground truth box G_r^{ℓ} in the image ℓ, *i.e*,

$$YP_{d,t,j}^{\ell} = \begin{cases} 1 \text{ if } \exists\ G_r^{\ell},\ r = 1, \ldots, M^{\ell}, \text{ such that } p_{d,t,j}^{\ell} \in G_r^{\ell}, \\ 0 \ \text{ otherwise.} \end{cases}$$

$\mathcal{L}_{calP,j}$ may pose a complexity issue as $|\mathcal{L}_{calP,j}| = \sum_{\ell=1}^{L} \sum_{t=1}^{N_j^{\ell}} \left|B_{t,j}^{\ell}\right|$. To avoid this, one may use a smaller set $\mathcal{L}'_{calP,j} \subset \mathcal{L}_{calP,j}$, which represents roughly the same information as $\mathcal{L}_{calP,j}$ and built as follows: for each triple (ℓ, t, j), only 10 couples among the couples $(S_{t,j}^{\ell}, YP_{d,t,j}^{\ell})$, $d = 1, \ldots, \left|B_{t,j}^{\ell}\right|$, are selected such that the ratio $\frac{\left|\{YP^{\ell}_{d,t,j}|d=1,\ldots,\left|B^{\ell}_{t,j}\right|,YP^{\ell}_{d,t,j}=1\}\right|}{\left|\{YP^{\ell}_{d,t,j}|d=1,\ldots,\left|B^{\ell}_{t,j}\right|,YP^{\ell}_{d,t,j}=0\}\right|}$ is preserved. $\mathcal{L}'_{calP,j}$ has then a size of $\left|\mathcal{L}'_{calP,j}\right| = 10\sum_{\ell=1}^{L} N^{\ell}_j$.

Set $\mathcal{L}_{calP,j}$ is useful for pixels that have a score, *i.e.*, are contained in a box. A pixel $p_{x,y}$ that does no belong to any box returned by a given detector D_j, does not have a score for this detector. Yet, it is reasonable to assume that D_j is almost certain that this pixel does not belong to a face, which can be modelled by a MF denoted $m^{\mathcal{P}_{x,y}}_{*,j}$. A first possibility for $m^{\mathcal{P}_{x,y}}_{*,j}$ could be to simply choose some MF representing this kind of knowledge, but this is not a very satisfying solution. Moreover, it should be taken into account that detectors do not present the exact same performances (*e.g.*, some may have many more pixels not in boxes than others). Within this scope, we propose a solution to obtain $m^{\mathcal{P}_{x,y}}_{*,j}$. For each detector D_j, its classification performance on pixels that do not belong to boxes is estimated using L images, where the positions of the faces really present are known. We denote by TN (True Negative) the number of pixels correctly classified on these images as non-face and FN (False Negative) the number of pixels classified as non-face but actually belonging to a face. $m^{\mathcal{P}_{x,y}}_{*,j}$ can then be defined by $m^{\mathcal{P}_{x,y}}_{*,j}(\{0\}) = \frac{TN}{TN+FN+1}, m^{\mathcal{P}_{x,y}}_{*,j}(\{1\}) = \frac{FN}{TN+FN+1}, m^{\mathcal{P}_{x,y}}_{*,j}(\{0,1\}) = \frac{1}{TN+FN+1}$. This definition may be seen as an evidential binning calibration [12] applied to pixels that do not belong to any of the boxes.

Our modeling of box absence is quite different from that of the box-based method, and arguably more consistent. Indeed, in this latter method, for a given area in an image, there are two different modelings of box absence for a detector depending on the situation: either none of the detectors has provided a box, in which case the area is considered as non face, which amounts to considering that the detectors know that there is no face; or only a subset of the detectors has provided a box, in which case the other detectors are ignored, which is equivalent (under Dempster's rule) to considering that these detectors know nothing.

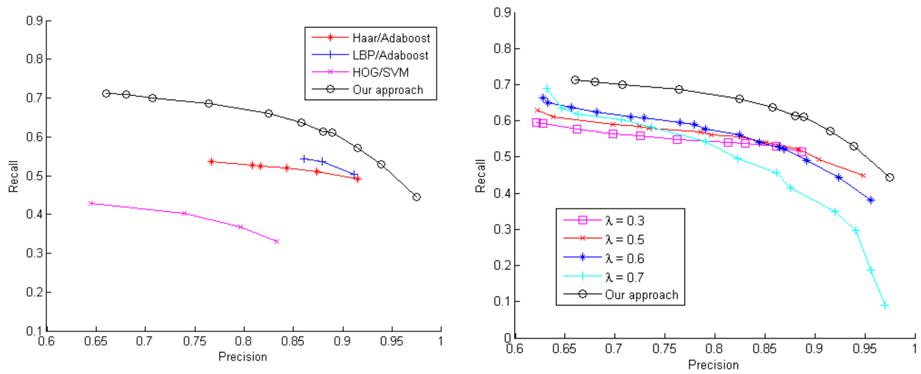
5 Experiment

In this section, the results of the proposed approach on a literature dataset are presented and compared to that of the box-based method presented in Section 3.

We selected three face detectors in the light of the availability of an open source implementation. The first detector is the one proposed by Viola and Jones [10], which is based on a classification algorithm called Adaboost and that uses Haar feature extraction. The second detector is a variant of the previous one: the same classification algorithm is used but with Local Binary Patterns (LBP) feature extraction [4]. The third detector relies on Support Vector Machine (SVM) and uses Histogram of Oriented Gradients (HOG) features [6].

For our experiment, we used a literature database called Face Detection Data Set and Benchmark (FDDB) [5]. It contains the annotations (ground truth) for 5171 faces in a set of 2845 images. In this paper, about 2000 images are used for the training of the detectors, and around 200 for calibration. Performance tests are conducted over the last 600 images, containing 1062 ground truth faces.

As our purpose is to minimize the number of non-blurred face pixels, it is worse to consider a face pixel as non-face than the opposite. In other words, using the decision strategy relying on upper expected costs (Section 2.1), decisions were taken for each test pixel with costs such that $c(1,0) <= c(0,1)$. More specifically, we fixed $c(1,0) = 1$ and gradually increased $c(0,1)$ starting from $c(0,1) = 1$, to obtain different performance points. To quantify performances, we used recall (proportion of pixels correctly blurred among the pixels to be blurred) and precision (proportion of pixels correctly blurred among blurred pixels).



(a) Versus detectors.

(b) Versus box-based method.

Fig. 1: Pixel-based approach vs detectors (1a) and vs box-based approach (1b).

Figure (1a) compare the results of the three selected detectors taken alone with our approach relying on a combination of their outputs. Comparison between the box-based approach used with different values of the overlap threshold λ and our approach is shown in Figure (1b).

6 Conclusion

In this paper, a pixel-based face blurring system relying on evidential calibration and fusion of several detector outputs was proposed. It brings several advantages over a previous box-based proposal: avoidance of the overlap threshold, of a clustering step, more consistent treatment of box absence, better performances. Several improvements are envisioned such as adding a skin colour detector to the system and refining the calibration and fusion steps. Some experiments are also envisaged on a more challenging database, which presents difficulties such as image quality or low light conditions.

References

1. T. Denoeux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.
2. M. Everingham, L. Van Gool, C. KI Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
3. F. Faux and F. Luthon. Theory of evidence for face detection and tracking. *International Journal of Approximate Reasoning*, 53(5):728–746, 2012.
4. A. Hadid, M. Pietikäinen, and T. Ahonen. A discriminative feature space for detecting and recognizing faces. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:II–797, 2004.
5. V. Jain and E. Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical Report UM-CS-2010-009, U Massachusetts, 2010.
6. E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–136, 1997.
7. J.C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. in large margin classifiers*, 10(3):61–74, 1999.
8. G. Shafer. *A mathematical theory of evidence*. Princeton University Press, 1976.
9. P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
10. P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.
11. P. Xu, F. Davoine, and T. Denœux. Evidential combination of pedestrian detectors. In *British Machine Vision Conference*, 2014.
12. P. Xu, F. Davoine, H. Zha, and T. Denoeux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning*, 72:55–70, 2016.